Jurnal Dinamika Informatika
Volume 11, No 2, September 2022
ISSN 1978-1660 : 103-112
ISSN *online* 2549-8517

Penerapan Algoritma Decision Tree Untuk Memprediksi Kanker Payudara menggunakan Data Mining dan Machine Learning

Ismail Setiawan[1], Erna Kumalasari Nurnawati [2]
[1]Prodi Sistem dan Teknologi Informasi, FST Universitas Aisyiyah Surakarta
[2]Prodi Informatika FTIB Institut Sains & Teknologi Akprind
e-mail: [1]ismail@aiska-university.ac.id, [2]ernakumala@akprind.ac.id

Intisari

Kanker payudara menjadi pembunuh nomor 1 dunia bahkan di Indonesia. Penangan penyakit ini umumnya dilakukan setelah penderita memasuki tahap kanker yang sudah lama bersemayam dalam tubuh. Penelitian ini mencoba untuk membuat model prediksi kanker payudara yang akurasinya diatas 99%. Harapannya adalah dengan model tersebut pendeteksian dini kanker payudara dapat dilakukan dengan diagosa yang sangat akurat. Sehingga membantu dokter atau fasilitas Kesehatan memberikan penangan sedini mungkian agar kanker payudara tidak berkembang. Model yang dikomparasikan adalah algoritma decision tree yaitu ID3, CART dan C4.5. Pengunaan pruning dan pre-pruning dilakukan untuk melihat akurasi model yang dibangun, hasilnya 98,25% didapat pada algoritma ID3 dan CART baik menggunakan atau tidak menggunakan pruning dan pre-pruning. Masih belum tercapai nilai akurasi 99%, kendala ini mungkin karena beberapa parameter yang belum ditemukan nilai yang pas.

Kata kunci— Akurasi, CART, C4.5, Kanker Payudara, ID3

Abstract

Even in Indonesia, breast cancer is the leading cause of death. Treatment for this illness typically begins when the patient reaches the stage of cancer that has been present in the body for a long time. The goal of this work is to develop a breast cancer prediction model that is above 99% accurate. It is hoped that this approach will enable highly accurate early identification of breast cancer. In order

to prevent breast cancer from developing, this aids medical professionals or healthcare institutions in providing treatment as soon as possible. The decision tree algorithm, specifically ID3, CART, and C4.5, is the model being compared. The accuracy of the model that was developed was tested using pruning and pre-pruning; the results were 98.25% for the ID3 and CART algorithms, whether or not these techniques were used. The fact that the accuracy level of 99% has not yet been reached could be caused by a number of parameters that have not yet found the proper value.

Keywords— Accuracy, CART, C4.5, Breast Cancer, ID3

PENDAHULUAN

Kanker payudara merupakan penyumbang kematian sebesar 6,6% dari seluruh kematian akibat kanker di dunia,dengan jumlah kasus sebesar 11,6% dari seluruh jenis kanker, sedangkan insiden kanker payudara pada perempuan di Indonesia sebesar 11,3% [1]. Secara nasional prevalensi penyakit kanker pada penduduk semua umur di Indonesia tahun 2013 sebesar 1,4% atau diperkirakan sekitar 347.792 orang. DI Yogyakarta memiliki prevalensi tertinggi untuk penyakit kanker, yaitu sebesar 4,1%, sedangkan prevalensi kanker payudara untuk propinsi Jambi sebesar 1,5% atau diperkirakan sekitar 4.995 penduduk [2]. Kanker payudara telah menjadi pembunuh nomor 1 di Indonesia bahkan di dunia [3]. Kanker ini menyerang Wanita walaupun pria juga memiliki potensi terkena kanker payudara [4][5][6]. Jumlah penduduk Wanita di Indonesia adalah 49,52% atau sekitar 136.361.271 jiwa dari 275.361.267 jiwa penduduk Indonesia. Jumlah Wanita di Indonesia yang terkena kanker payudara per tahun 2022 sebanyak 60.234 jiwa. Beberapa diantaranya datang ke fasilitas kesehatan dengan kondisi yang stadium lanjut [7]. Jika datang lebih awal kemungkinan indakan pencegahan dapat di lakukan sehingga mampu menolong nyawa penderita[8].

Pemerintah menanggapi serius hal tersebut, melalui Kementrian Kesehatan (kemenkes) dilakukan Gerakan yang dinamakan SADARI (Sadar sendiri) [9]. Gerakan tersebut diharapkan mampu menekan angka penderita kanker payudara stadium lanjut dengan pencegahan sejak dini. Gerakan tersebut perlu di

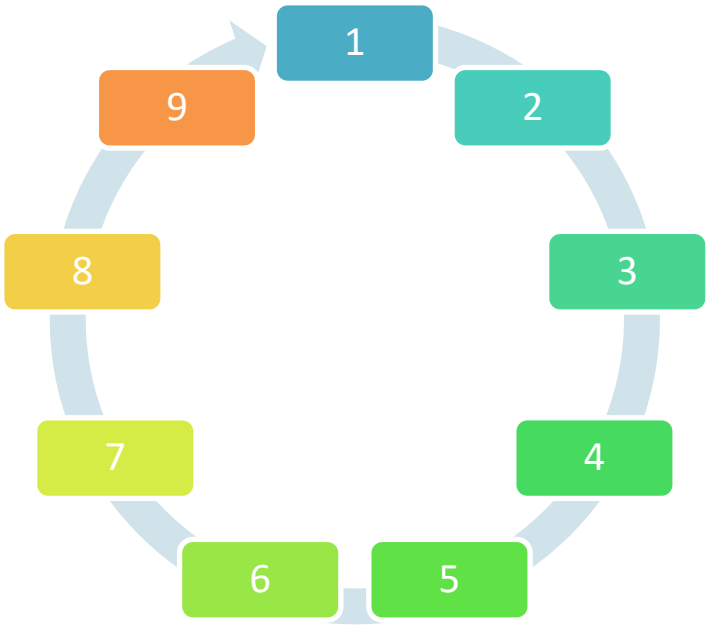
Jurnal Dinamika Informatika
Volume 11, No 2, September 2022
ISSN 1978-1660 : 103-112
ISSN *online* 2549-8517

sosialisasikan keseluruh orang [10], karena terbukti meningkatkan kesadaran dan Langkah antisipasi orang terhadap kanker payudara. Kanker payudara stadium lanjut dapat dikelompokan dalam kanker jinak atau ganas [11][12]. Kanker jinak masih dapat dilakukan prosedur perawatan yang terukur dan terjangkau untuk mengatasinya. Namun jika sudah masuk kanker ganas maka kemoterapi dan perawatan khusus perlu di terapkan[13][14].

Pandemic covid-19 menerapkan pembatasan sosial, sehingga sosialisasi dari kemenkes menjadi terbatas ruang geraknya [15]. Namun demikian interaksi yang dilakukan secara daring (dalam jaringan) meningkat drastis [16]. Media promosi Kesehatan menjadi beragam dan sangat mudah di akses dari mana saja dan kapan saja [17].

METODE PENELITIAN

Penelitian ini menggunakan metode crips-DM dengan improve pada beberapa langkah.



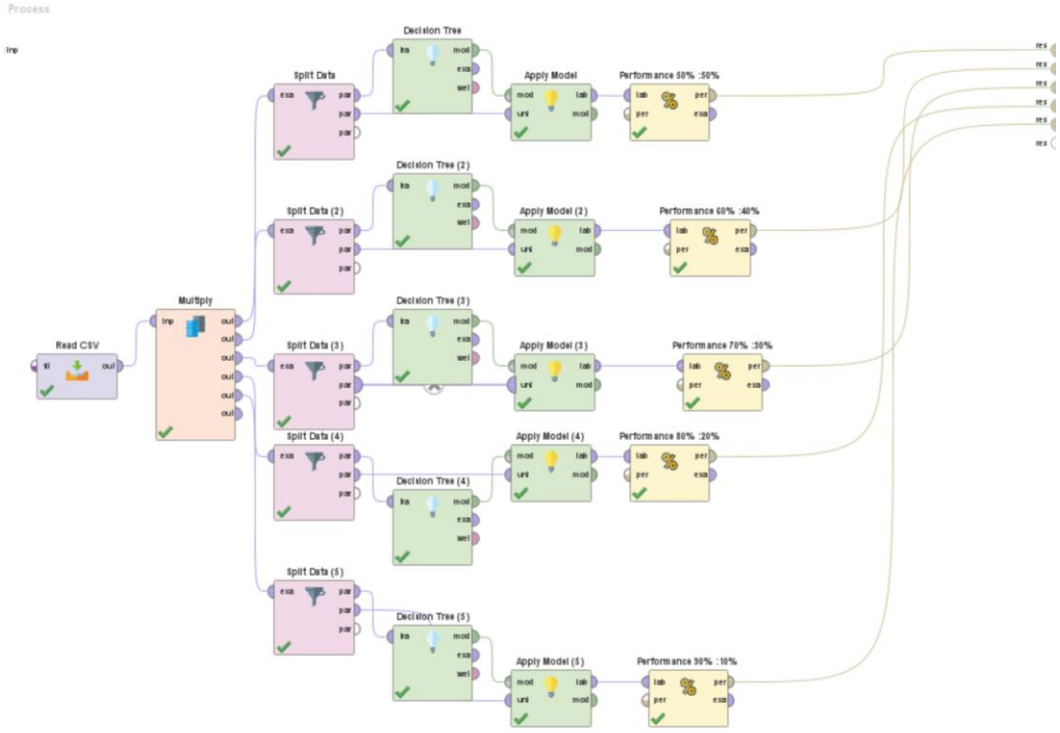
Gambar1. Metode penelitian

1. Merubah data menjadi numerik
2. Menghapus nilai yang hilang
3. Menghilangkan data yang duplikat
4. Melakukan normalisasi data (range transformation)

5. Memilih atribut yang menjadi target lalu mencari korelasi dengan atribut yang lain dengan nilai positif paling tinggi.
6. Memisahkan data menjadi data testing dan data training dengan perbandingan 50% : 50%, 60% : 40%, 70% : 30%, 80% : 20%, 90% : 10%.
7. Mencari performa model yang dibuat.
8. Perubahan parameter untuk mendapatkan performa yang paling tinggi.

HASIL DAN PEMBAHASAN

Tahap awal adalah membangun model dengan tool machine learning untuk mempercepat proses perhitungan kalkulasi. Penelitian ini menggunakan rapid miner sebagai tool nya.



Gambar 2 Pemodelan Yang Dilakukan Menggunakan Rapid Miner

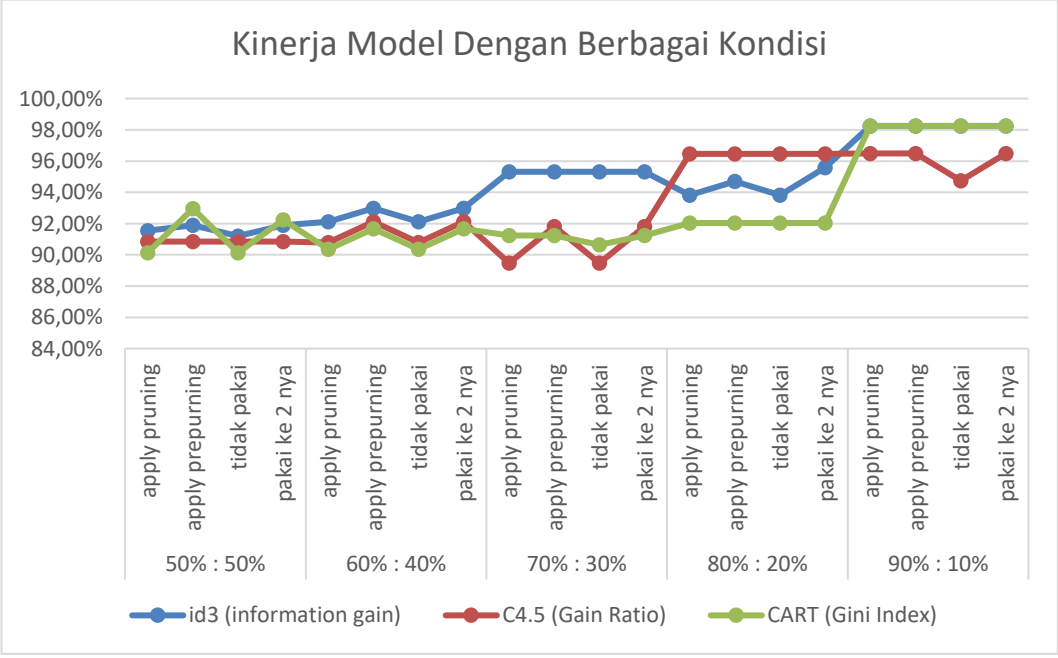
Operator pertama yang digunakan adalah read csv, opertaror ini digunakan untuk membaca data csv tentang kanker payudara yang sudah tersedia di UCI dataset. Penggunaan data public dipilih agar penelitian ini bisa di duplikasi dan dilanjutkan untuk penelitian kedepan. Selanjutnya operator multiply dipakai untuk mempercepat mengetahui hasil dari masing masing keadaan jika data di split atau dipisah dengan ketentuan 50% : 50%, 60% : 40%, 70% : 30%, 80% : 20%, 90% :

Jurnal Dinamika Informatika
Volume 11, No 2, September 2022
ISSN 1978-1660 : 103-112
ISSN *online* 2549-8517

10%. Kemudian operator decision tree digunakan untuk melihat kinerja masing masing model yang ada yaitu ID3, CHART dan C4.5 [18].

Pada operator decision tree terdapat pilihan penggunaan apply pruning dan apply pre-pruning. Purning adalah mengidentifikasikan dan membuang cabang yang tidak diperlukan pada pohon yang telah terbentuk [19]. Hal ini dikarenakan pohon keputusan yang dikontruksi dapat berukuran besar, maka dapat disederhanakan dengan melakukan pemangkasan berdasarkan nilai kepercayaan (confident level)[20]. Sedangkan pre-pruning adalah menghentikan pembangunan suatu subtree lebih awal (dengan memutuskan untuk tidak lebih jauh mempartisi data training) [21]. Saat seketika berhenti, maka node berubah menjadi leaf (node akhir). Node akhir ini menjadi kelas yang paling sering muncul di antara subset sampel [8]. Proses uji pada penelitian ini dilakukan dengan menggunakan atau tidak menggunakan dari ke dua pilihan tersebut. Hasil kinerja dapat dilihat pada tabel 1.

Selanjutnya operator apply model digunakan untuk menerapak hasil split data dan operator decision tree. Untuk melihat kinerja digunakan operator performance. Pada Operator performance nilai yang ingin dilihat pada penelitian ini adalah akurasinya. Parameter-parameter lain pada operator performance sementara tidak digunakan. Tujuan dari penelitian ini adalah mencari model yang menghasilkan akurasi diatas 99%.



Gambar 3 Kinerja Model Dengan Berbagai Kondisi

Tabel 1 Hasil Kinerja Komparasi Model

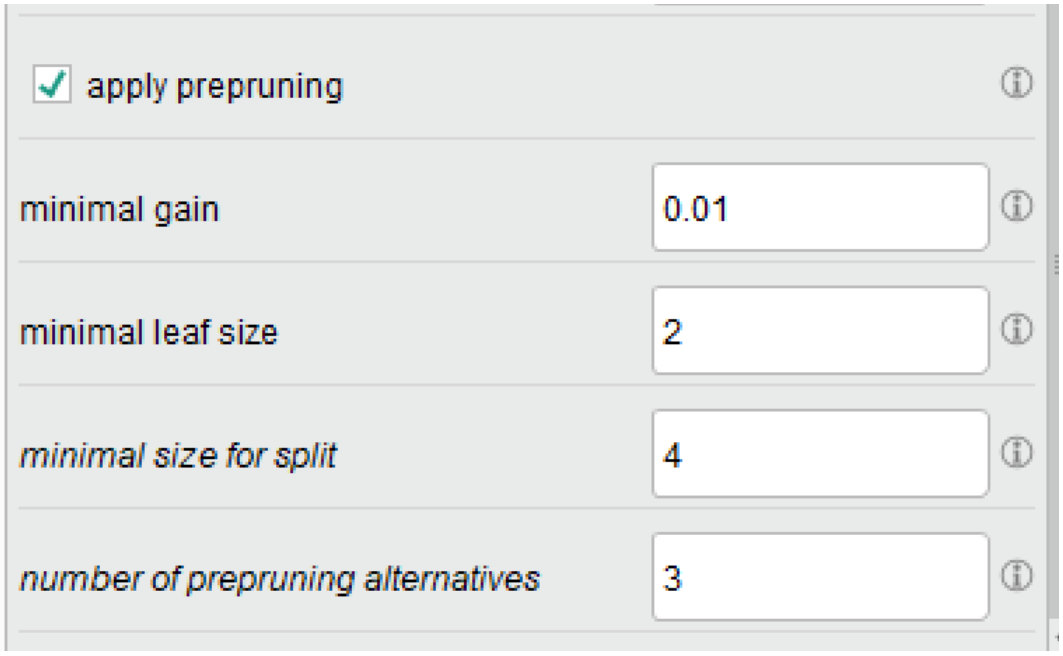
SPLIT DATA	MODEL	ID3	C4.5	CART
50% : 50%	Apply Pruning	91,55%	90,85%	90,14%
	Apply Prepurning	91,90%	90,85%	92,96%
	Tidak Pakai	91,20%	90,85%	90,14%
	Pakai Ke 2 Nya	91,90%	90,85%	92,25%
60% : 40%	Apply Pruning	92,11%	90,79%	90,35%
	Apply Prepurning	92,98%	92,11%	91,67%
	Tidak Pakai	92,11%	90,79%	90,35%
	Pakai Ke 2 Nya	92,98%	92,11%	91,67%
70% : 30%	Apply Pruning	95,32%	89,47%	91,23%
	Apply Prepurning	95,32%	91,81%	91,23%
	Tidak Pakai	95,32%	89,47%	90,64%
	Pakai Ke 2 Nya	95,32%	91,81%	91,23%
80% : 20%	Apply Pruning	93,81%	96,46%	92,04%
	Apply Prepurning	94,69%	96,46%	92,04%
	Tidak Pakai	93,81%	96,46%	92,04%
	Pakai Ke 2 Nya	95,58%	96,46%	92,04%
90% : 10%	Apply Pruning	98,25%	96,49%	98,25%
	Apply Prepurning	98,25%	96,49%	98,25%
	Tidak Pakai	98,25%	94,74%	98,25%
	Pakai Ke 2 Nya	98,25%	96,49%	98,25%

KESIMPULAN

Melihat hasil kinerja model berdasarkan split data, penggunaan 90% data training dan 10% testing menunjukan nilai yang paling tinggi. Model ID3 dan CART selalu menunjukan akurasi yang sama disetiap kondisi antar menggunakan atau tidak menggunakan pruning dan pre-pruning yaitu 98,25%. Namun demikian tujuan dari penelitian ini adalah menghasilkan akurasi diatas 99% belum tercapai. Masih ada beberapa parameter yang dapat dirubah sehingga kemungkinan akurasi daiatas 99% bisa dicapai. Pada pruning ada nilai confidence yang dapat di rubah angkanya seperti gambar 3 begitupun pre-purning seperti gambar 4.

Jurnal Dinamika Informatika
Volume 11, No 2, September 2022
ISSN 1978-1660 : 103-112
ISSN *online* 2549-8517

Gambar 4. Parameter pruning



Gambar 5. Parameter pre-pruning

SARAN

Penelitian ini masih belum mencapai angka akurasi diatas 99%. Pasalnya dalam dunia Kesehatan nilai akurasi sebuah model sebelum diterapkan sebaiknya diatas 99% karena berkaitan dengan keselamatan jiwa seseorang. Oleh karena itu perlu peningkatan nilai yang dapat dicapai dengan merubah parameter-parameter yang ada pada masing masing model. Kegiatan tersebut dapat dilakukan kedepan berdasarkan apa yang telah dilakukan pada penelitian ini.

DAFTAR PUSTAKA

[1] C. de Martel, D. Georges, F. Bray, J. Ferlay, and G. M. Clifford, "Global burden of cancer attributable to infections in 2018: a worldwide incidence analysis," *Lancet Glob. Heal.*, vol. 8, no. 2, pp. e180–e190, 2020.

[2] R. I. Kemenkes, "Pusat data dan informasi kementerian kesehatan RI: Situasi dan analisis diabetes," *Di akses pada*, vol. 8, 2019.

[3] E. Marfianti, "Peningkatan Pengetahuan Kanker Payudara dan Ketrampilan Periksa Payudara Sendiri (SADARI) untuk Deteksi Dini Kanker Payudara di Semutan Jatimulyo Dlingo," *J. Abdimas Madani dan Lestari*, pp. 25–31, 2021.

[4] A. N. Giaquinto *et al.*, "Breast cancer statistics, 2022," *CA. Cancer J. Clin.*, vol. 72, no. 6, pp. 524–541, 2022.

[5] Y. Lestari, "Sosialisasi sadari sebagai langkah awal pencegahan kanker payudara pada remaja putri sma sekabupaten sumbawa," *J. Pengabdi. Kpd. Masy. Indones.*, vol. 2, no. 2, pp. 180–185, 2022.

[6] H. E. Situmorang, H. Lumbantobing, Y. Kristina, K. Suweni, and D. A. Nurfaizah, "Pelatihan Deteksi Dini Kanker Payudara Dengan Metode 'SADARI'(Periksa Payudara Sendiri) pada Siswi-Siswi Sma Teruna Bakti di Jayapura Papua," *J. Kreat. Pengabdi. Kpd. Masy.*, vol. 5, no. 7, pp. 2152–2159, 2022.

[7] A. H. Narzulloyevich, M. G. Fazliddinovna, and K. F. Sharopovna, "Comparison of the results of modern methods of treatment of elderly women with breast cancer," *Eurasian Med. Res. Period.*, vol. 3, pp. 9–15, 2021.

[8] A. Hassan Zadeh, Q. Alsabi, J. E. Ramirez-Vick, and N. Nosoudi, "Characterizing basal-like triple negative breast cancer using gene expression analysis: A data mining approach," *Expert Syst. Appl.*, vol. 148, p. 113253, 2020, doi: https://doi.org/10.1016/j.eswa.2020.113253.

[9] S. M. Hutagaol, "Tingkat Pengetahuan, Sikap, Dan Perilaku Mahasiswi

Tentang Pemeriksaan Payudara Sendiri (Sadari) Di Universitas Sumatera Utara Tahun 2020," 2021.

[10] J. Kusumawaty, E. Noviati, I. Sukmawati, Y. Srinayanti, and Y. Rahayu, "Efektivitas Edukasi SADARI (Pemeriksaan Payudara Sendiri) Untuk Deteksi Dini Kanker Payudara," *ABDIMAS J. Pengabdi. Masy.*, vol. 4, no. 1, pp. 496–501, 2021.

[11] S. B. Akben, "Determination of the Blood, Hormone and Obesity Value Ranges that Indicate the Breast Cancer, Using Data Mining Based Expert System," *IRBM*, vol. 40, no. 6, pp. 355–360, 2019, doi: https://doi.org/10.1016/j.irbm.2019.05.007.

[12] P. Heudel *et al.*, "1427P - Analysis of prognostic factors on overall survival in elderly women treated for early breast cancer using data mining and machine learning," *Ann. Oncol.*, vol. 30, pp. v580–v581, 2019, doi: https://doi.org/10.1093/annonc/mdz257.022.

[13] S. Simsek, U. Kursuncu, E. Kibis, M. AnisAbdellatif, and A. Dag, "A hybrid data mining approach for identifying the temporal effects of variables associated with breast cancer survival," *Expert Syst. Appl.*, vol. 139, p. 112863, 2020, doi: https://doi.org/10.1016/j.eswa.2019.112863.

[14] T. Jaikuna, M. Aznar, P. Hoskin, M. Van Herk, C. West, and E. Vasquez Osorio, "PO-1822 Feasibility of spatial normalisation for image-based data mining in breast cancer radiotherapy," *Radiother. Oncol.*, vol. 161, pp. S1552–S1553, 2021, doi: https://doi.org/10.1016/S0167-8140(21)08273-6.

[15] O. Alagoz *et al.*, "Impact of the COVID-19 pandemic on breast cancer mortality in the US: estimates from collaborative simulation modeling," *JNCI J. Natl. Cancer Inst.*, vol. 113, no. 11, pp. 1484–1494, 2021.

[16] S. Guo *et al.*, "Micro-tomographic and infrared spectral data mining for breast cancer diagnosis," *Opt. Lasers Eng.*, vol. 160, p. 107305, 2023.

[17] T. O. Nielsen *et al.*, "Assessment of Ki67 in breast cancer: updated recommendations from the international Ki67 in breast cancer working group," *JNCI J. Natl. Cancer Inst.*, vol. 113, no. 7, pp. 808–819, 2021.

[18] M. Fabregue, S. Bringay, P. Poncelet, M. Teisseire, and B. Orsetti, "Mining

microarray data to predict the histological grade of a breast cancer," *J. Biomed. Inform.*, vol. 44, pp. S12–S16, 2011, doi: https://doi.org/10.1016/j.jbi.2011.03.002.

[19] M. Ture, F. Tokatli, and I. Kurt Omurlu, "The comparisons of prognostic indexes using data mining techniques and Cox regression analysis in the breast cancer data," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 8247–8254, 2009, doi: https://doi.org/10.1016/j.eswa.2008.10.014.

[20] A. Ansaripour, K. Zendehdel, C. A. Uyl - de Groot, A. NaemiSanatdost, and W. K. Redekop, "PCN59 - Direct Medical Costs Of Her2 Positive Breast Cancer Management In Iran: A Claims Database And Data Mining Analysis," *Value Heal.*, vol. 18, no. 3, pp. A199–A200, 2015, doi: https://doi.org/10.1016/j.jval.2015.03.1156.

[21] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," *Artif. Intell. Med.*, vol. 34, no. 2, pp. 113–127, 2005, doi: https://doi.org/10.1016/j.artmed.2004.07.002.